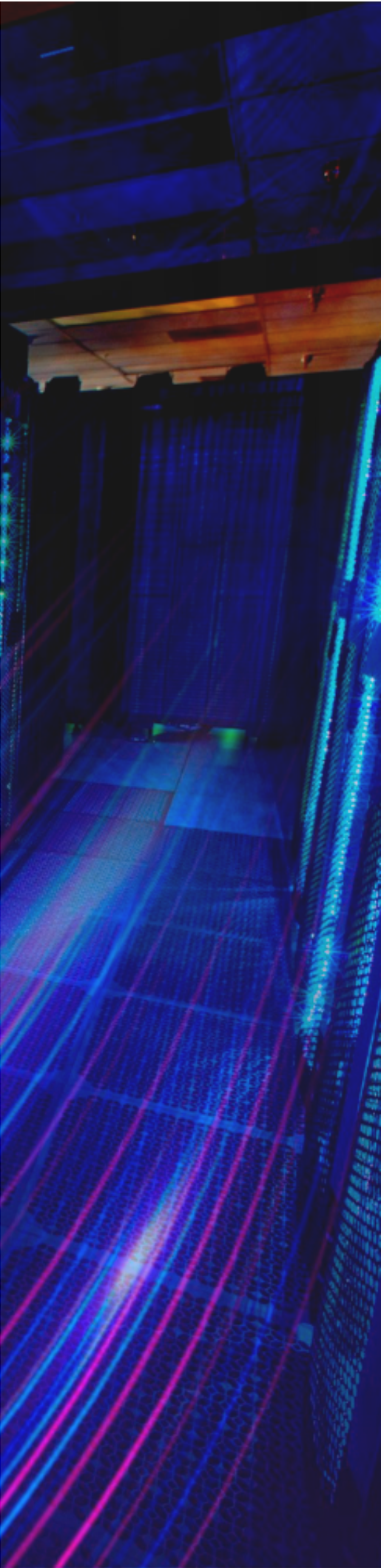
LA TIENDA DE DATOS

Para IA

ASSIST

Gold
Business
Partner

IBM

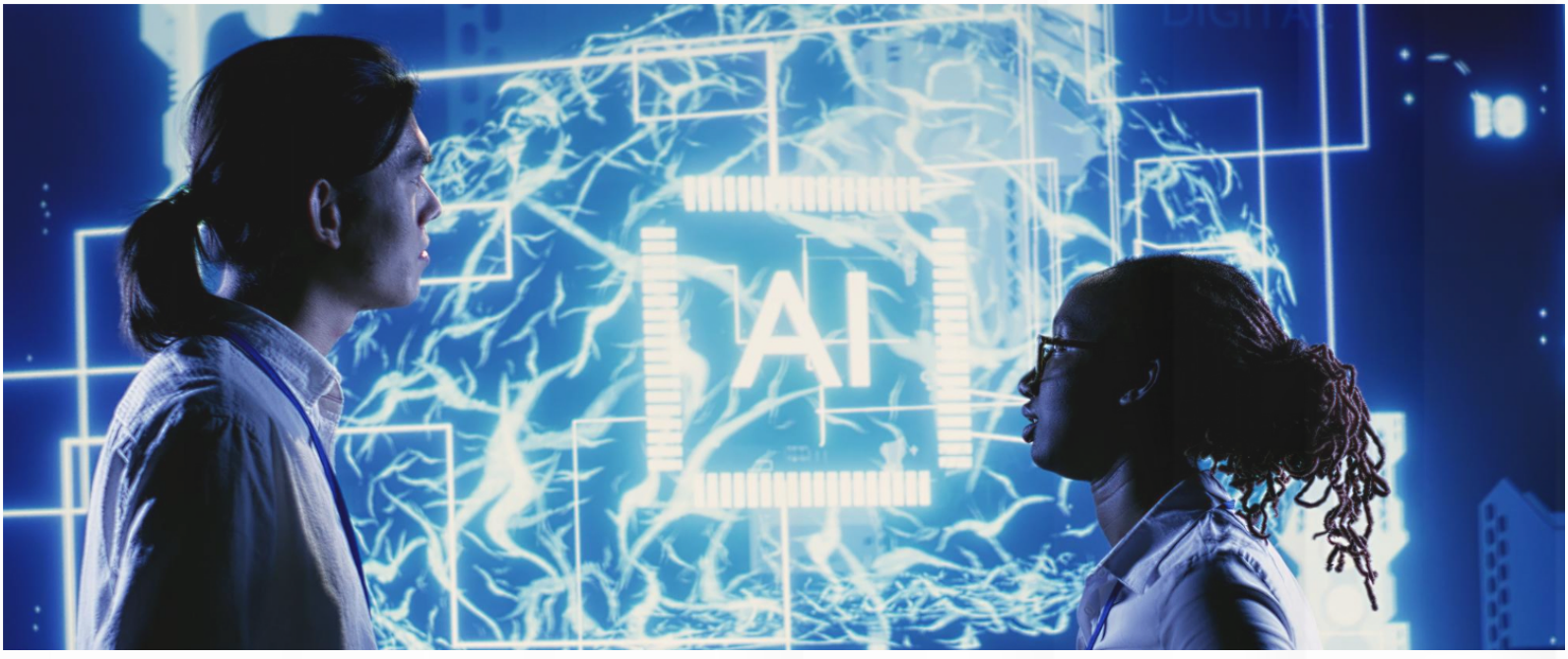


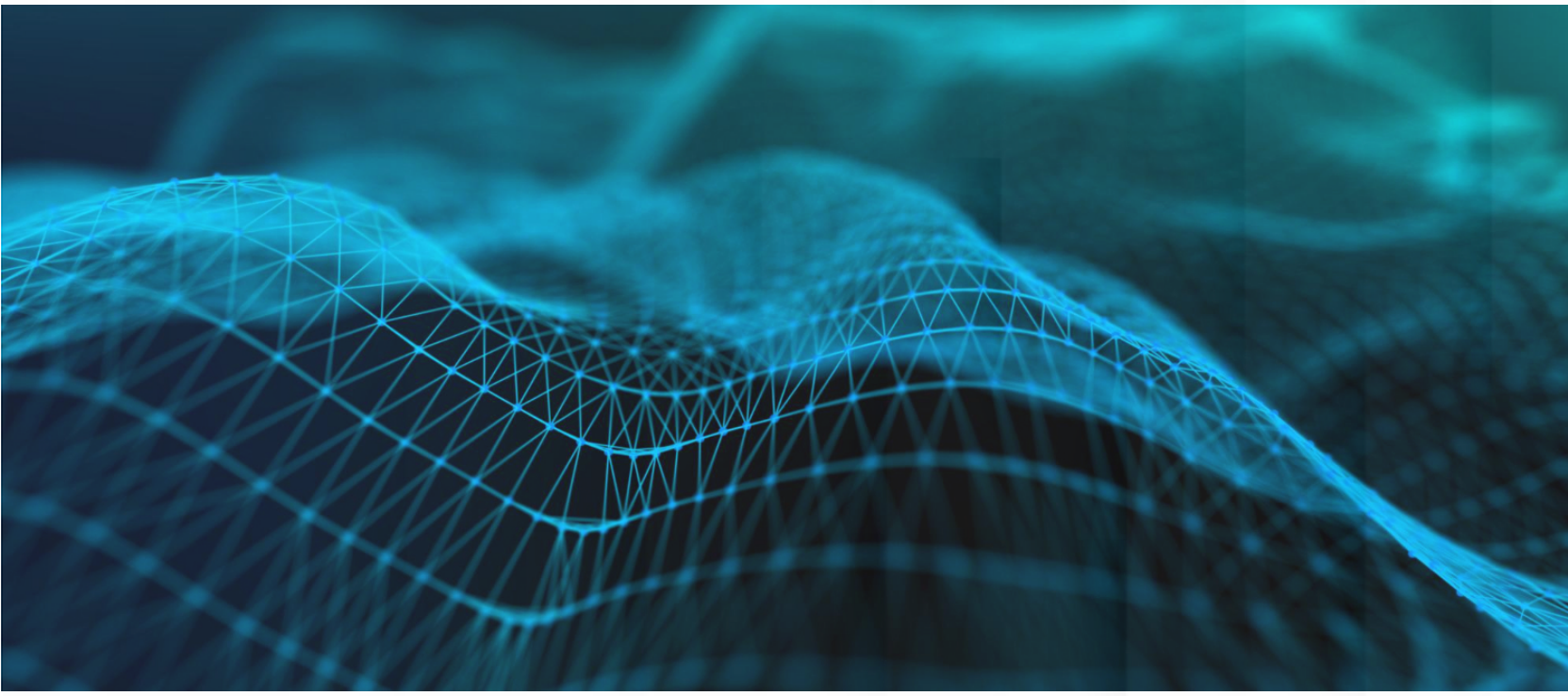
CONTENIDO

INTRODUCCION

Este ebook examinará la última solución de gestión de datos abiertos para líderes de datos y análisis que desean reducir significativamente costos, simplificar el acceso a datos y automatizar la gobernanza unificada para escalar la IA. Es hora de la data lakehouse. Los datos están en el centro de cada negocio. Mantienen el funcionamiento de las aplicaciones, potencian las perspectivas predictivas y permiten mejores experiencias para clientes y empleados.

Pero el beneficio completo de los datos es esquivo debido a la forma en que se almacenan y se accede a ellos para el análisis y la IA. No estás solo si dependes de repositorios monolíticos con múltiples almacenes de datos y lagos de datos, en las instalaciones y en la nube; el 82% de las organizaciones están inhibidas por los silos de datos. Y está a punto de empeorar: según IDC, se espera que la cantidad de datos almacenados crezca un 250% para 2025.





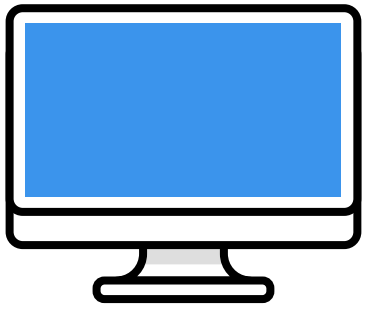
El datalake se suponía que solucionaría todos estos problemas; solo hay que depositar los datos en un lugar centralizado y procesarlos.
Pero no es tan fácil actualizar los lagos, catalogar correctamente los datos o garantizar una buena gobernanza, y los conjuntos de habilidades requeridos para estas tareas son específicos, raros y costosos.
Como resultado, los lagos de datos han resultado costosos de construir y mantener.

Un almacén de datos ofrece un alto rendimiento para procesar terabytes de datos estructurados. Pero los almacenes también pueden volverse costosos, especialmente para cargas de trabajo nuevas y en evolución. La mayoría de las organizaciones ejecutan cargas de trabajo de análisis e IA en ecosistemas que son complejos y costosos.

Es hora de un cambio.

Se espera que la cantidad de datos almacenados crezca un

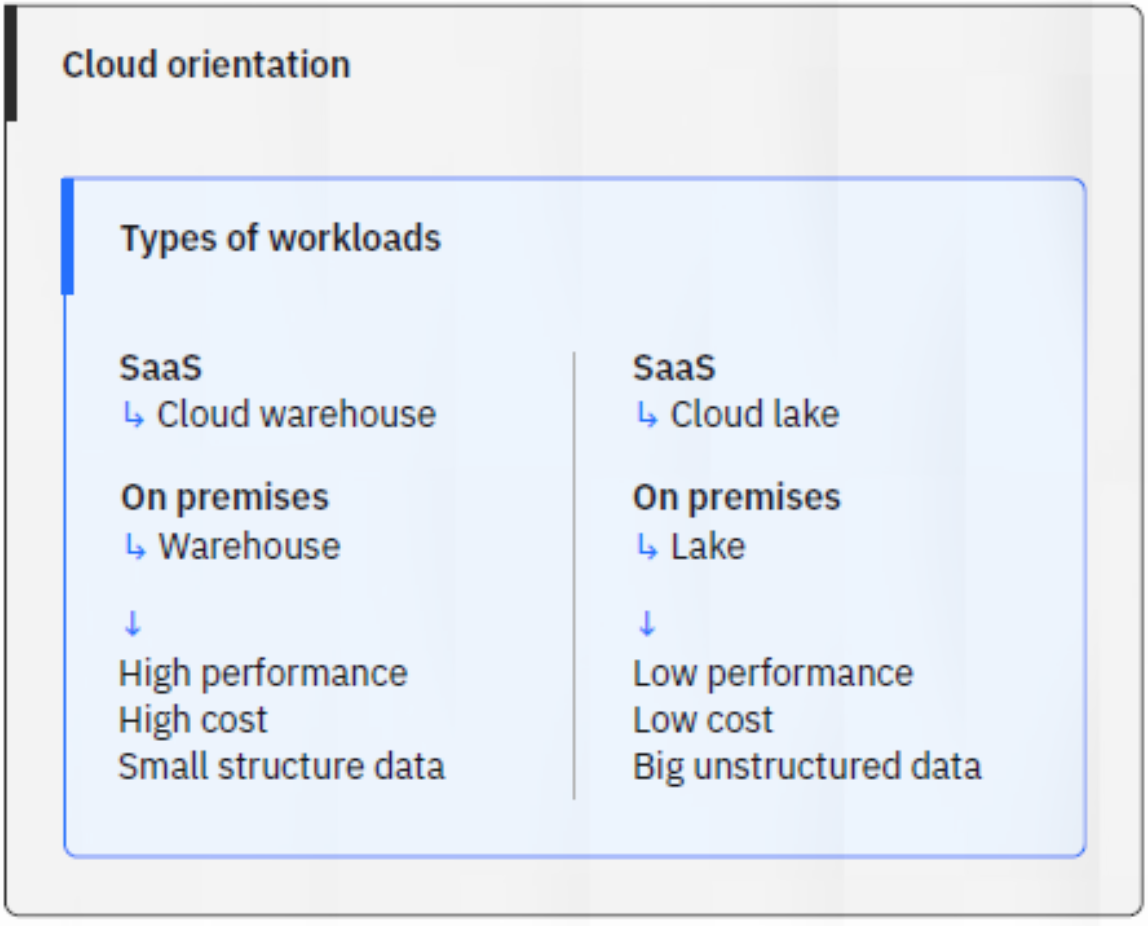
250% para 2025



EL ESTADO ACTUAL DE LA ARQUITECTURA DE DATOS

2

Una combinación de almacenes en las instalaciones y nativos de la nube y lagos de datos personalizados es común en la arquitectura empresarial hoy en día. Es probable que encuentres que equilibrar costos, datos en silos y gobernanza de datos son desafíos constantes.

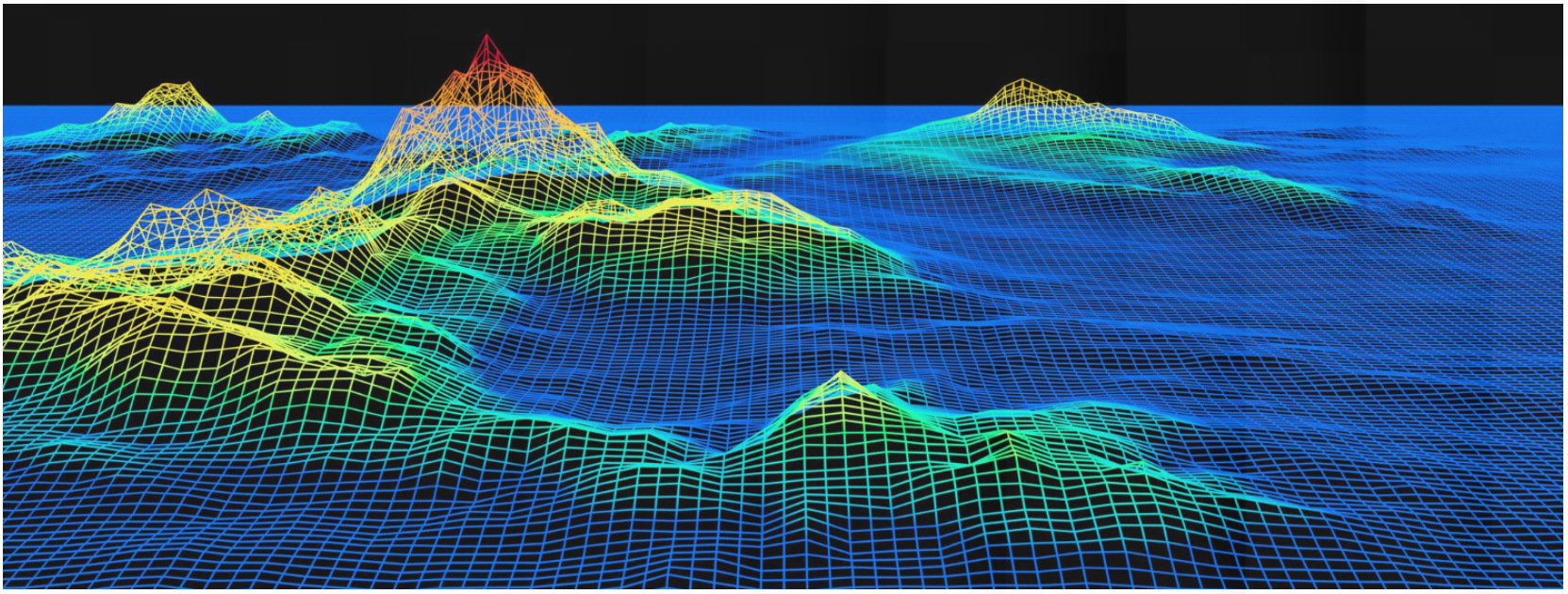


Datalake es un cambio de paradigma emergente en cómo las empresas obtienen conocimientos

DATA LAKEHOUSE DEFINIDA

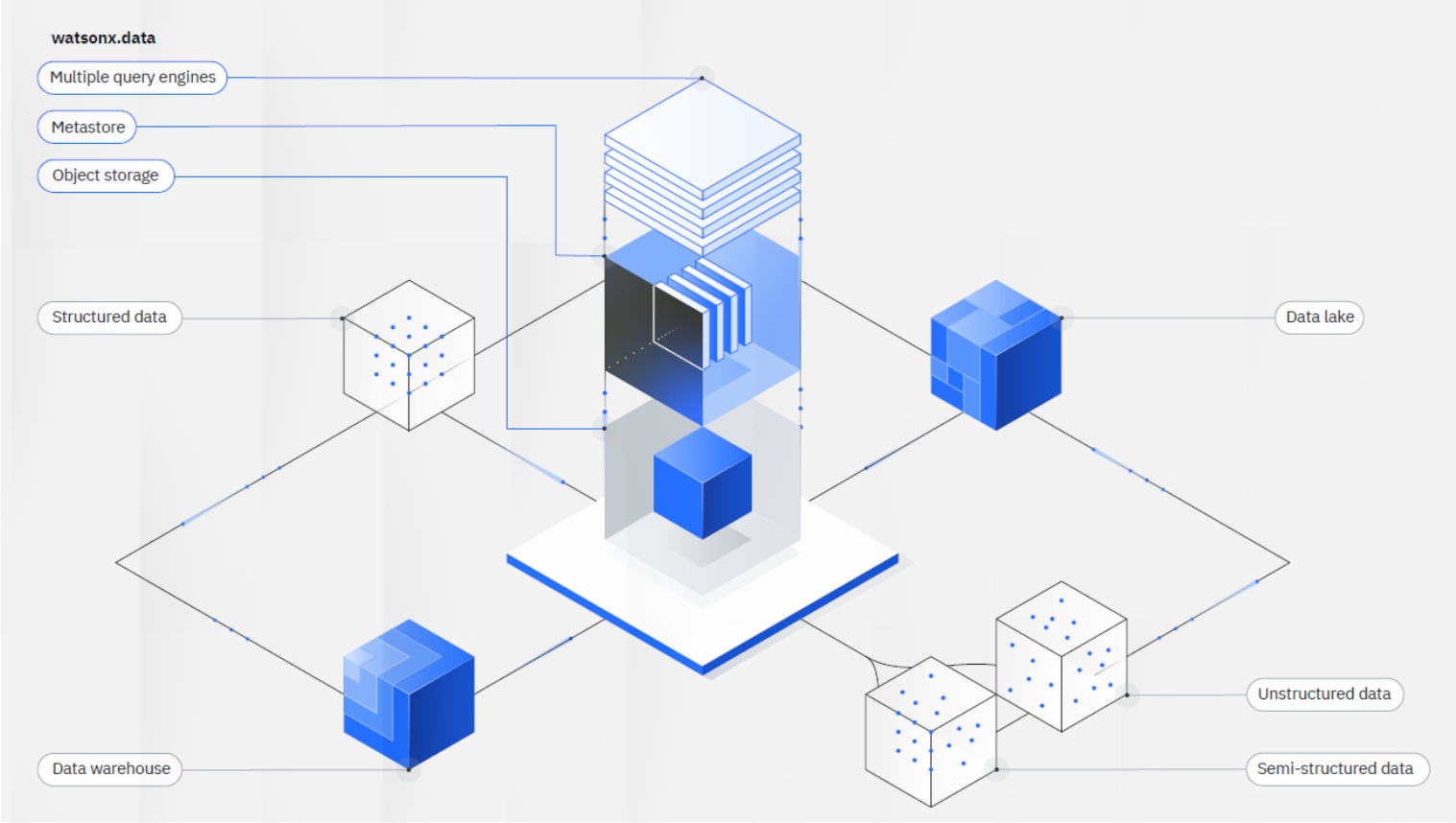
3

Busca una solución de lakehouse que proporcione una base de datos moderna para escalar la inteligencia artificial.



La Data Lakehouse busca una solución que proporcione una base de datos moderna para escalar la IA. La data lakehouse de datos es una arquitectura emergente que ofrece la flexibilidad de un lago de datos con el rendimiento y la estructura de un almacén de datos.

La mayoría de las soluciones de data lakehouse ofrecen un motor de consulta de alto rendimiento sobre un almacenamiento de bajo costo en conjunto con una capa de gobernanza de metadatos. Las capas de metadatos inteligentes facilitan a los usuarios categorizar y clasificar datos no estructurados, como video y voz, y datos semi-estructurados, como XML, JSON y correos electrónicos.



La mejor data lakehouse ofrecerá tecnologías de código abierto que reducen la duplicación de datos y simplifican los complejos flujos de trabajo de ETL. Ten en cuenta que algunas data lakehouse de primera generación tienen limitaciones clave que limitan su capacidad para abordar los desafíos de costos y complejidad.

Por ejemplo, un único motor de consulta diseñado para cargas de trabajo de inteligencia empresarial o aprendizaje automático (ML) podría ser ineficaz cuando se utiliza para otro tipo de carga de trabajo. El equipo de datos e IA de IBM cree que cada carga de trabajo es única y debe optimizarse con el entorno más adecuado que mantenga los costos al mínimo y el rendimiento al máximo. Elige una data lakehouse que ofrezca un nivel óptimo de rendimiento para una mejor toma de decisiones, junto con la flexibilidad necesaria para desbloquear valor de todos los tipos de datos.

COMPONENTES DE LA ARQUITECTURA

4

Infraestructura

Este componente es donde se implementará tu lakehouse, totalmente gestionada en cualquier nube o entorno local.

Almacenamiento

Esta capa es donde los datos están físicamente almacenados, los cuales se guardan como archivos y pueden ser almacenados en formatos de datos abiertos, como Apache Parquet y Avro. Los formatos de datos abiertos son especificaciones de archivos y protocolos puestos a disposición de la comunidad de código abierto para que cualquiera pueda ingresar y mejorarlos.

Formatos de tabla abiertos

Formatos de tabla abiertos, como Apache Iceberg, te ayudan a proporcionar estructura y ofrecer la confiabilidad y simplicidad de SQL con datos a gran escala. Estos formatos permiten que diferentes motores accedan a los mismos datos al mismo tiempo, lo que ayuda a evitar el bloqueo de proveedores.
Compartir datos entre múltiples herramientas y repositorios de datos, como tu almacén de datos; una única copia de datos te permite reducir la duplicación de datos y romper los silos.

Gobernanza

También se almacena metadatos con formatos de tabla abiertos; sirve para definir los formatos de archivo para cualquier herramienta que pueda leer o escribir formatos de datos abiertos.

Servicio de metadatos

Este componente es necesario para comprender qué datos están disponibles en la capa de almacenamiento. El motor de consulta requiere los metadatos para los datos y tablas para proporcionar un linaje completo y saber dónde están ubicados, cómo lucen y cómo leerlos.

Catálogos de datos

Este componente ayuda a los usuarios a encontrar los datos correctos para la tarea y proporciona información semántica para políticas y reglas. Espera almacenar metadatos comerciales como terminologías comerciales y etiquetas para habilitar búsqueda y protección de datos.

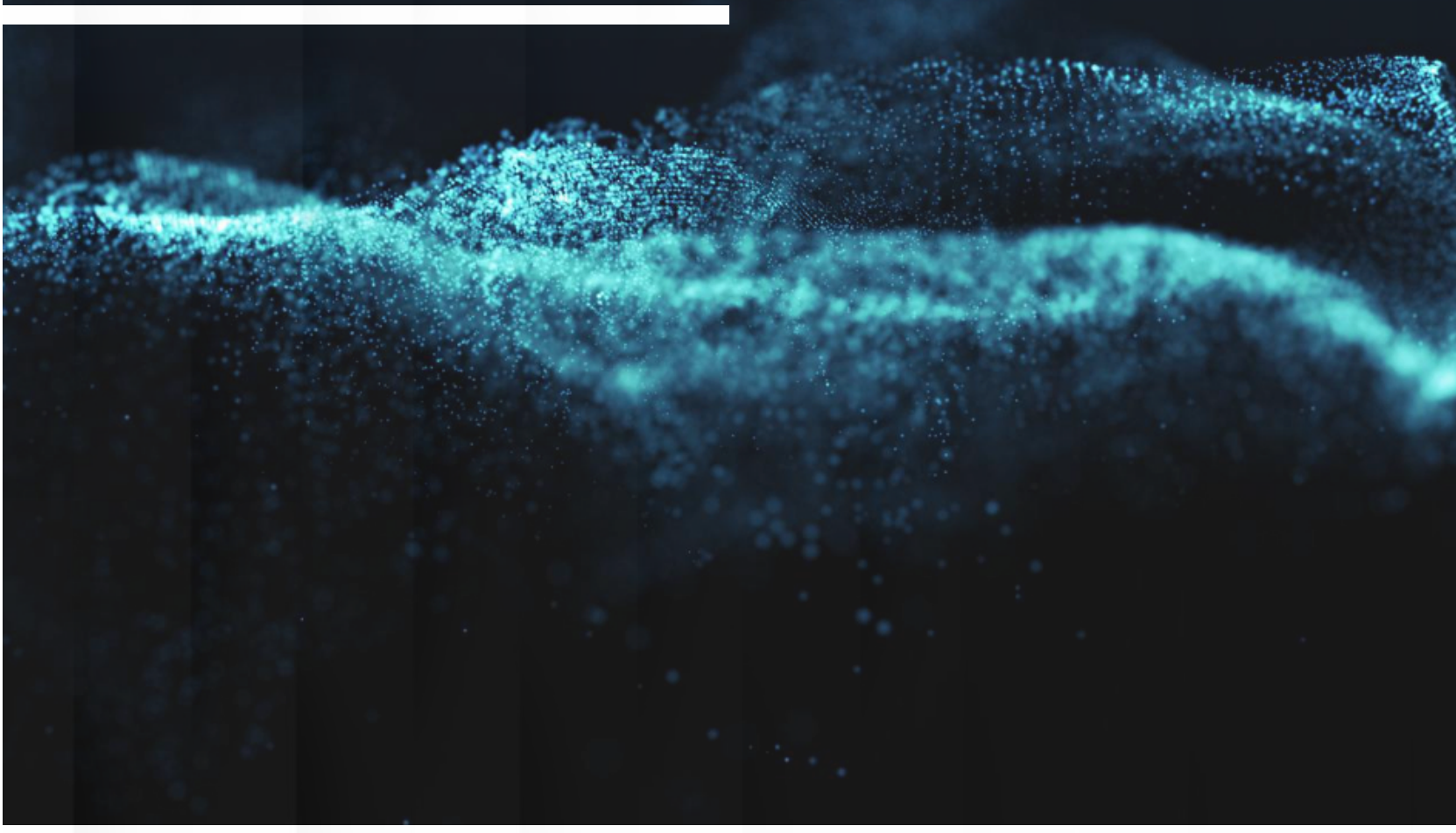
Motor de políticas

Este componente permite a los usuarios definir políticas de protección de datos y habilita al motor para hacer cumplir esas políticas. Para crear un marco de gobierno escalable, a menudo se despliega un motor de políticas con el servicio de metadatos técnico y el catálogo de datos.

Motor de consulta

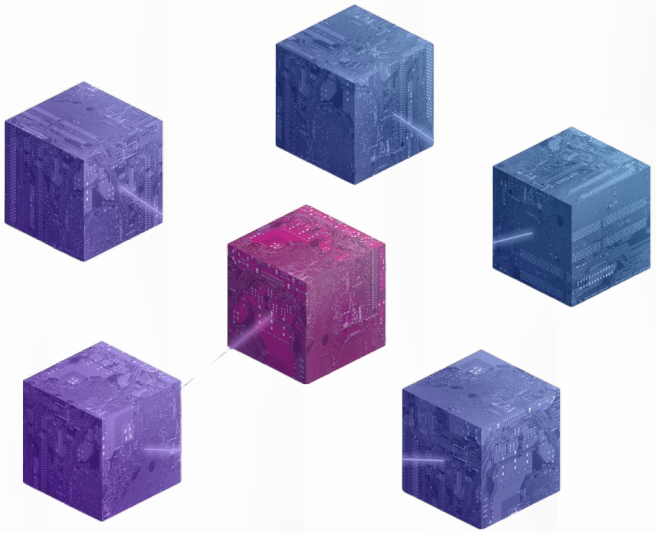
Este componente está en el corazón de la data lakehouse abierto. Un motor de consulta, que puede ser de código abierto o propietario, accede a los datos en formato de tabla abierta y a menudo se conoce como el componente de cómputo. Los motores de consulta típicamente vienen en dos tipos: un motor de consulta basado en SQL, como el de código abierto Presto, o un motor de Apache Spark de código abierto o su equivalente.

En una arquitectura de lakehouse abierta, el motor de consulta es completamente modular, lo que significa que el motor puede ser escalado dinámicamente para satisfacer las demandas de carga de trabajo y concurrencia. Los motores de consulta también pueden adjuntarse a cualquier catálogo y almacenamiento.



50%

Ahora es posible obtener insights más rápidos y confiables mientras reduces los costos del almacén de datos a la mitad.

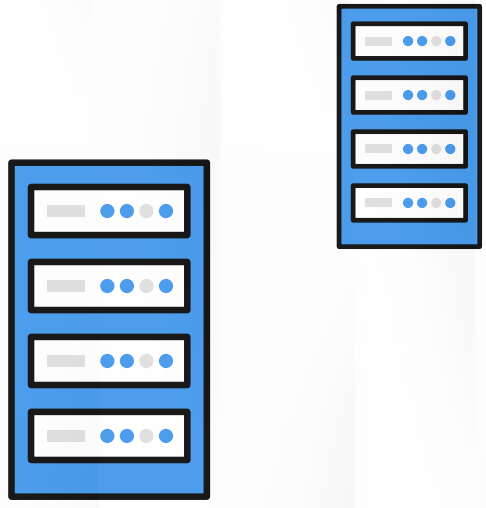


OPORTUNIDADES DE OPTIMIZACIÓN DE COSTOS 5

Si tu organización tiene implementaciones existentes de big data en las instalaciones, un lakehouse ofrece una alternativa menos costosa para almacenar datos en formatos abiertos en almacenamiento de objetos. Reducirás el costo de análisis, disminuirás la complejidad y mejorarás el tiempo de valoración. Si tienes una implementación de almacén existente, un enfoque de lakehouse puede representar una alternativa de costo más bajo y escalable masivamente para tus cargas de trabajo de análisis grandes que son menos sensibles a los acuerdos de nivel de servicio (SLAs).

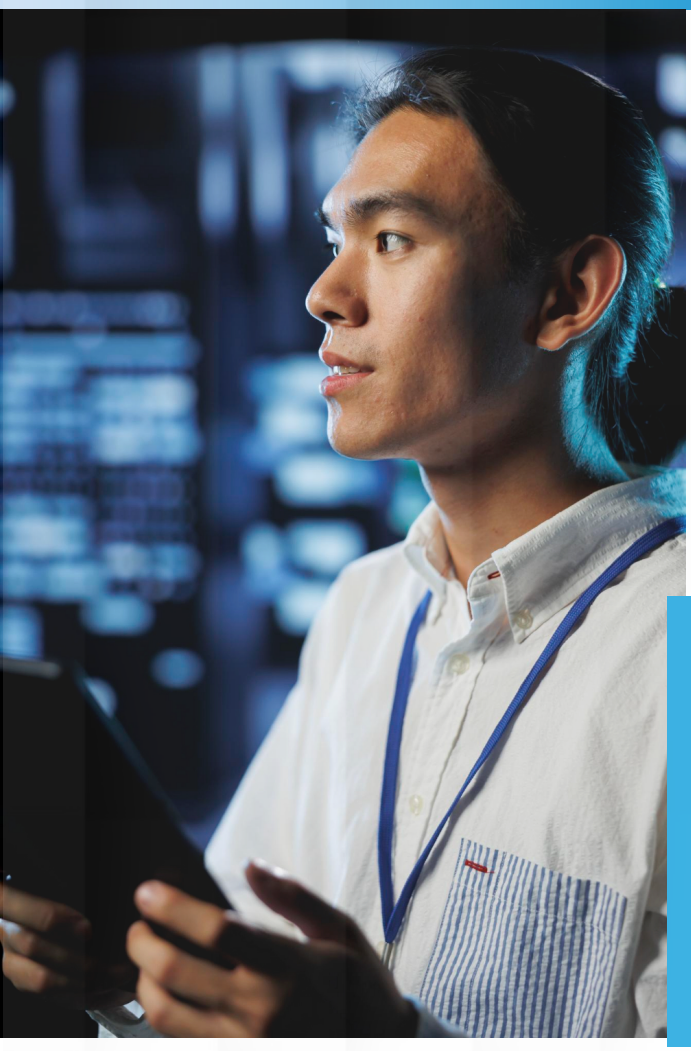
Los almacenes suelen ser costosos y propietarios, pero con un lakehouse, puedes reducir drásticamente los costos de almacenamiento y cómputo. Puedes optimizar las cargas de trabajo del almacén utilizando motores adecuados para el propósito que se basen en los requisitos de tu carga de trabajo. La naturaleza abierta de un lakehouse te libera de la tecnología de almacén propietaria, lo que significa menos bloqueo de proveedores y una reducción en los costos generales de infraestructura de TI.

IBM WatsonX.data es un almacén de datos abierto, híbrido y gobernado optimizado para todas las cargas de trabajo de datos, análisis e Inteligencia Artificial.



MEJORAS EN ANÁLISIS Y CIENCIA DE DATOS

6



"Nos estamos moviendo en la dirección donde el lakehouse de datos se convierte en una mejor práctica."

Adam Ronthal
Vice President
Gartner

Los formatos de datos propietarios y los altos costos de almacenamiento limitan la colaboración y despliegue de modelos de IA y ML dentro de un entorno de almacén de datos; los data lakes se enfrentan a cargas de trabajo de ciencia de datos de bajo rendimiento. El aislamiento de estas tecnologías ha llevado a desafíos de infraestructura aguas abajo, junto con las implicaciones de seguridad y gobierno que surgen de la duplicación y movimiento de datos para el desarrollo de modelos de IA y ML.

Un data lakehouse es una excelente manera de ayudar a los colegas que están ávidos por los insights que esperan en los datos de tu organización. Si estás comprometido en extraer valor empresarial del torrente de datos que se acerca a ti, considera seriamente la estrategia del lakehouse.

Adam Ronthal, vicepresidente y analista en Gartner, afirma que ***"Nos estamos moviendo en la dirección donde el lakehouse de datos se convierte en una mejor práctica".*** La mejor aproximación ofrecerá un entorno abierto, colaborativo y gobernado para la gestión

Examinemos IBM® WatsonX.Data™: el almacén de datos abierto, híbrido y gobernado optimizado para todas las cargas de trabajo de datos, análisis e inteligencia artificial.

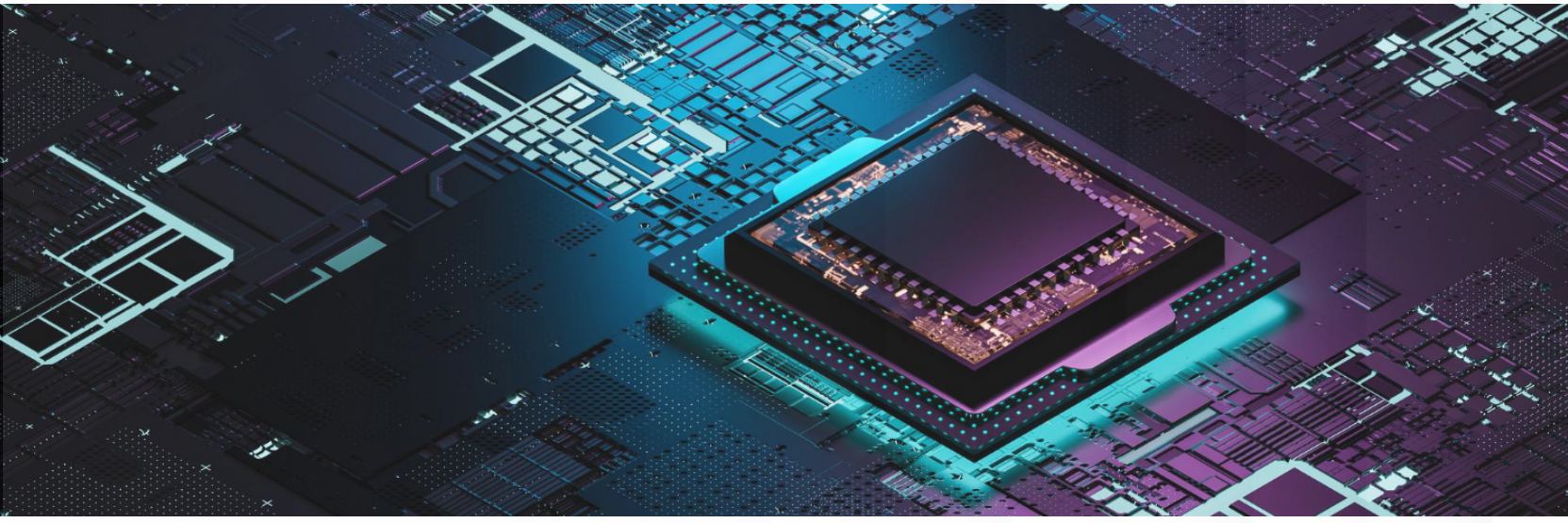
IBM watsonx.data

Escalar cargas de trabajo de inteligencia artificial para todos tus datos, en cualquier lugar. Watsonx.data es un almacén de datos abierto, híbrido y regulado, optimizado para todos los datos, análisis y cargas de trabajo de inteligencia artificial, construido sobre una arquitectura de lago de datos **(ver figura 1).**

Accede a todos tus datos y maximiza la cobertura de cargas de trabajo en todos tus entornos híbridos en la nube. Espera una implementación sin problemas de un servicio totalmente gestionado en cualquier nube o entorno local. Accede a cualquier fuente de datos, donde sea que resida, a través de un único punto de entrada y combínalo utilizando formatos de datos abiertos. Intégralo en tu entorno existente con código y estándares abiertos, y con interoperabilidad con servicios de IBM y de terceros.

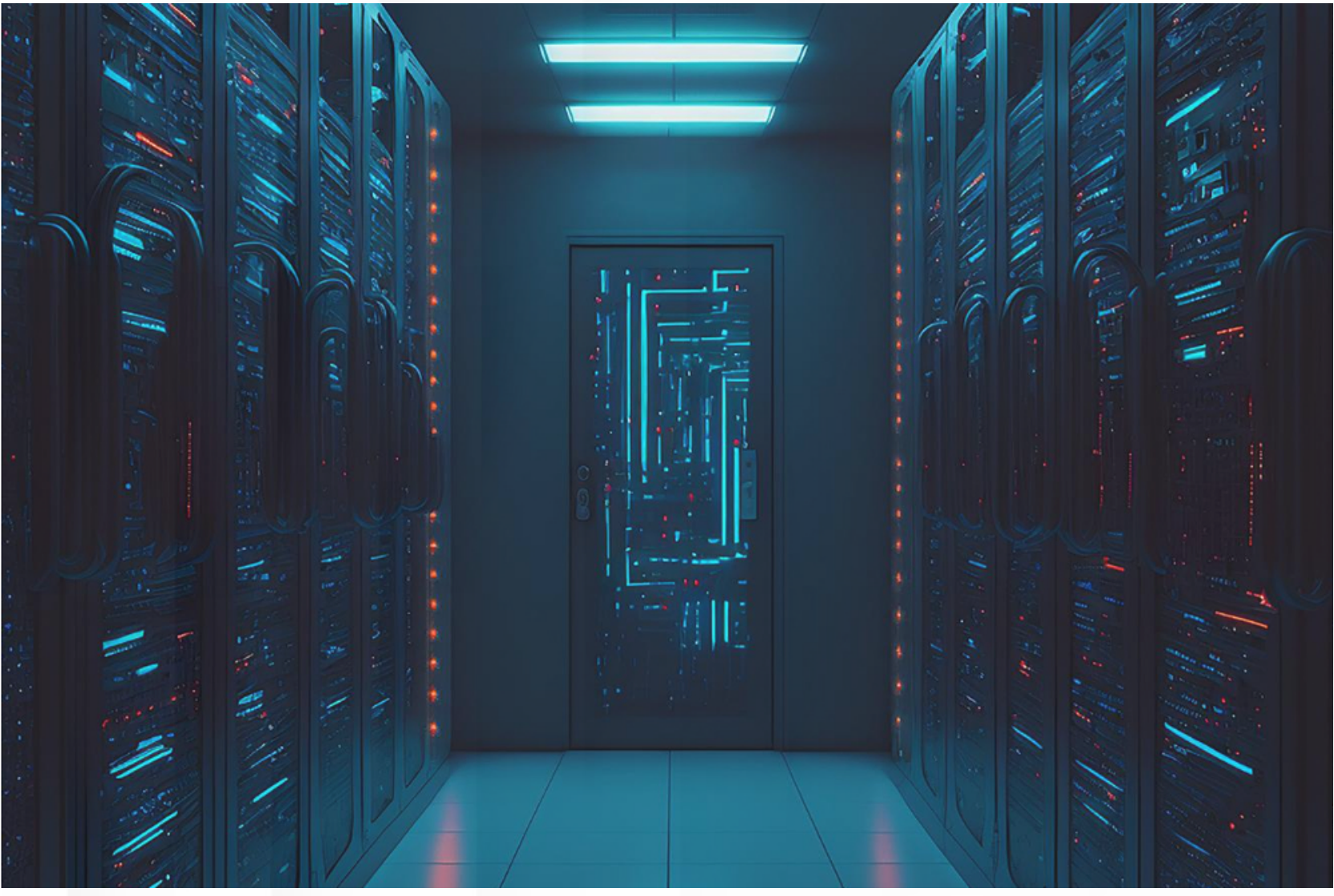
Acelera el tiempo para obtener ideas confiables. Comienza rápidamente con gobierno y automatización integrados; fortalece el cumplimiento empresarial y la seguridad con un gobierno unificado en todo tu ecosistema. Una interfaz de usuario clara y una consola de clic y listo ayudan a tus equipos a ingresar, acceder y transformar datos y ejecutar cargas de trabajo. Observa lo rápido que adoptarán un panel de control que les facilite ahorrar dinero y ofrecer ideas frescas y confiables.

Reduce el costo de tu almacén de datos hasta en un 50% mediante la optimización de cargas de trabajo en múltiples motores de consulta y niveles de almacenamiento. Optimiza las costosas cargas de trabajo del almacén de datos con motores adecuados que escalan automáticamente hacia arriba y hacia abajo. Reduce los costos eliminando la duplicación de datos cuando utilices almacenamiento de objetos de bajo costo; extrae más valor de los datos en lagos de datos poco efectivos.



Siguientes pasos

Aprovecha el conocimiento en gestión y optimización de datos del equipo de **IBM junto a Assist,** perfeccionado para enfrentar las cargas de trabajo de datos más exigentes del mundo.
Observa lo rápido que puedes obtener valor de Watsonx.data.



Contacto

comunicaciones@assist.com.co

Cra 62 No. 103 – 44 / Ofc 501

Torre del Reloj, Bogotá, Colombia

Tel: (601) 432-2360

Cel: 310-216-9640

ASSIST

Gold
Business
Partner

IBM